A magyar nyelv sajátságaihoz illeszkedő módszerek szövegek automatikus osztályozására

Németh András[1,2], Balázs László[1]

[1] Alkalmazott Logikai Laboratórium,
Hankóczy J. u. 7. 1022 Budapest,
{xandrew, bazsi}@all.hu
2 Budapesti Műszaki és Gazdaságtudományi Egyetem
Számítástudományi és Információelméleti Tanszék,
Magyar tudósok körútja 2. 1117 Budapest,
xandrew@cs.bme.hu

Kivonat: A magyar nyelv gazdag morfológiája és agglutináló jellege megkérdőjelezi az angol nyelvre jól működő szövegklasszifikációs technikák változatlan alkalmazását. A legtöbb bevett módszerben szavak előfordulását vizsgáljuk a dokumentumokban, azonban a magyar nyelv esetében a szóalakok nagy száma miatt ez nem tűnik alkalmas megközelítésnek. Jelen cikkben két módszert javaslunk a probléma kezelésére: a már korábban is alkalmazott szótövesítést, illetve n-grammok alapján történő osztályozást. Vizsgálataink azt mutatják, hogy a kisebb apparátust igénylő n-gramm alapú technikák is a szótövesítéshez hasonlóan jó eredményt adnak, és még robosztusabbnak is bizonyulnak annál.

1 Bevezetés

Az automatikus szövegosztályozási feladatban dokumentumokat kell előre meghatározott kategóriákba sorolnunk, adott és már kategóriákba sorolt mintadokumentumok alapján.

A feladat hosszú idő óta aktív kutatási terület, és a lehetséges megoldások iránt folyamatosan nő az érdeklődés az Interneten elérhető dokumentumok számának rohamos emelkedésével. Egy jól működő osztályozó problémák rendkívül széles körében használható, például keresési eredmények strukturált megjelenítésére, beérkező levelek automatikus szortírozására, vállalati intraneten rendezetlenül megtalálható dokumentumok elérhetővé tételére témák szerint tagoltan, IR (information retrieval) rendszerek fontos részeként és még hosszan sorolhatnánk.

Az igényeknek megfelelően a nemzetközi irodalom is hatalmas, módszerek széles skáláját javasolták a különböző sztohasztikus modellektől a döntési fákon és a legközelebbi szomszéd algoritmuson keresztül a neurális hálózatokig. Az egyik legsikeresebb megközelítésnek az SVM (support vector machine, lásd pl. [6]) osztályozó használata bizonyult.

A magyar szövegek osztályozásának specifikus feladata lényegesen kisebb figyelmet kapott. Kornai és társai az origo.hu portál kulcsszó kereső és téma osztályozó

rendszerének kiépítése kapcsán foglalkoztak a feladattal, és bemutattak egy Bayes-modell alapú osztályozót és szótövesítést használó megoldást [5].

Cavnar és Trenkle már nagyon korán javasolták nyelvfelismerésre és szövegklasszifikációra az n-grammok frekvencia profile-jainak összehasonlítását [1]. Később Langdon ajánlott 3-gramm előfordulási vektorokra alkalmazott legközelebbi szomszéd osztályozót nyelvfüggetlen dokumentumklasszifikációs technikaként [7].

Jelen cikkben áttekintjük az egyik legsikeresebb módszercsaládot, és megvizsgáljuk, hogy miért kevéssé alkalmasak a klasszikus technikák magyar nyelvű szövegek esetében (2. fejezet). Megadjuk a leírt algoritmus egy egyszerű általánosítását, és ezen általánosítás speciális eseteként ajánlunk két módszert a magyar nyelvhez illesztéshez: a szótövek ill. az n-grammok alapján történő osztályozást (3. fejezet). Ismertetjük a módszerek értékelésére elvégzett kísérleteinket (4. fejezet). Végül összefoglaljuk a kísérletekből levont következtetéseinket (5. fejezet).

2 Szó alapú szövegklasszifikációs módszerek áttekintése

Ebben a fejezetben áttekintünk egy bevett, jól működő osztályozási technikát, és megvizsgáljuk a magyar nyelvre való alkalmazhatóságát.

2.1 A feladat

Adott kategória címkékkel ellátott dokumentumok egy halmaza. A feladat ezen tanítókészlet alapján valamiféle modell betanulása, és ennek segítségével címkézetlen dokumentumok osztályozása. A szövegklasszifikációs feladat speciális esete a topic felismerés. Itt egy bizonyos kategóriába tartozó (pl. adott témáról szóló, innen az elnevezés) dokumentumokat kell elkülöníteni az összes többitől. A cikkben elvégzett kísérletek erre a speciális esetre vonatkoznak.

2.2 Dokumentum reprezentáció

Az elterjedt klasszifikációs módszerek közös eleme, hogy a dokumentumokról csak azt tartják nyilván, hogy mely szavak fordulnak bennük elő. Itt a „szó” jelentése előre definiált elválasztó karakterek (pl. whitespace, központozás) közötti összefüggő karaktersorozat. A sorrend információ minden esetben eldobódik, de általában még a pontos előfordulási számmal sem foglalkozunk, így egy dokumentumot a benne előforduló szavak halmazaként reprezentálhatunk (ez az úgynevezett bag of words model).

Bár ezzel a reprezentációval nyilvánvalóan rengeteg információt elveszítünk, mégis úgy tűnik – a jó minőségű elérhető klasszifikáció miatt –, hogy az osztályozási feladat szempontjából ez még megengedhető. Ráadásul az ezen modell alapján történő osztályozásra jól működő egyszerű módszereink vannak, míg a sorrendiségben rejlő plusz segítség felhasználása nehezen képzelhető el nagyon komoly apparátus, lényegében természetes nyelvű szövegértés nélkül.

2.3 Releváns szavak kiválogatása

Ezen technikák feladata, hogy az osztályozó lépés előtt csökkentsük a feladat dimenzióját a lehető legkevesebb információ elvesztése mellett. Lényegében az összes módszer azon alapul, hogy az egyes szavak kapnak egy pontszámot, ami jellemzi a relevanciájukat az adott osztályozási feladat szempontjából, és ezen függvény szerinti legfontosabb néhány szót tartjuk meg. A lehetséges függvények széles köre és értékelésük megtalálható [3]-ban. Mi az alábbi két, széles körben alkalmazott függvényt használtuk.

Khi négyzet (χ^2 , chi-square) függvény. Ez a statisztikában különféle hipotézisvizsgálatoknál használt próbán alapul. Lényegében függetlenség vizsgálatot végzünk: független-e a szó előfordulása a dokumentumok kategóriáitól. Egy kétdimenziós diszkrét mintasorozatból a két dimenzió függetlenségének megállapítására az alábbi statisztikát használhatjuk:

$$T = n\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(N_{ij} - \frac{N_{i.}N_{.j}}{n}\right)^2}{N_{i.}N_{.j}} \qquad (\mathbf{10})$$

Itt N_{ij} azon mintaelemek száma, ahol az első dimenzióban az *i*. a másodikban pedig a j. lehetséges érték vevődik fel. $N_{i.}$ ill. $N_{.j}$ a sor ill. oszlopösszegeket jelöli, *n* pedig a minták száma. Az így kapott T statisztika a nullhipotézis teljesülése esetén eloszlásban az $(r-1)(s-1)$ rendű χ^2 eloszláshoz tart, így minél nagyobb a *T*, annál inkább el kell vetni a függetlenséget a próba során, a konkrét kritikus érték a próba terjedelme alapján számítható. Ez alapján a fenti *T* jó eszköz a két dimenzió összefüggőségének mérésére. A mi esetünkben a két dimenziós mintasorozat első dimenziója a dokumentumok kategóriája, második dimenziója pedig 0 vagy 1 attól függően, hogy a vizsgált szó szerepel-e a megfelelő dokumentumban. Tehát az (1) függvény szerinti minél nagyobb értékű szavakat kell választani.

Az információs nyereség (information gain) függvény. Ennek meghatározásához először kiszámítjuk a kategóriák eloszlásának entrópiáját a teljes dokumentum halmazon. Ezután kiszámítjuk az entrópiát a vizsgált szót tartalmazó és nem tartalmazó dokumentumokra külön-külön is. Ennek a két entrópiának a halmazok (a szót tartalmazó dokumentumok halmaza ill. a szót nem tartalmazó dokumentumok halmaza) elemszámával súlyozott átlaga kisebb egyenlő az eredeti entrópiánál. Ha kisebb, az azt jelenti, hogy azzal, hogy tudjuk, hogy a szó szerepel-e egy dokumentumban információt nyerünk annak kategóriájáról. A két fenti érték különbsége az információs nyereség, itt is a minél magasabb információs nyereségű szavakat érdemes kiválasztani.

2.4 Osztályozás

Az irodalom feltűnően egységes abban a kérdésben, hogy milyen osztályozót érdemes használni dokumentum klasszifikációhoz. A lineáris SVM mutatja a legjobb

eredményeket. Ezért itt bővebben nem foglalkozunk az osztályozó választás kérdésével, végig lineáris SVM-et használunk.

2.5 A teljes modellépítési és értékelési folyamat áttekintése

A fenti módszer egy variánsának teszteléséhez a következő lépéssorozatot kell végrehajtanunk:

1. Korpusz választás ill. előállítás: Kiválasztjuk az osztályozandó kategóriákat, és előállítunk egy (dokumentum, kategória) párokból álló halmazt, amiben egy dokumentum csak egyszer szerepelhet, és minden dokumentum a vizsgált kategorizálás szerint helyesnek tartott kategóriájával áll párban.
2. A korpuszt két diszjunkt részre, tanító és tesztkészletre bontjuk szét.
3. Kiválasztunk egy relevancia függvényt, és a tanító készlet alapján minden szónak kiszámítjuk a relevanciáját, és kiválasztjuk az első k legrelevánsabb szót.
4. A kiválogatott szavak segítségével minden dokumentumhoz egy bitsorozatot rendelünk: minden bit egyes értéke azt jelenti, hogy a hozzá tartozó szó előfordul a szövegben. Az így kapott vektorok klasszifikációjára lineáris SVM-mel építünk osztályozó modellt.
5. A teljes osztályozó modell a 3. lépésben nyert szó listából és a 4. lépésben nyert SVM modellből áll.
6. Meghatározzuk a tesztkészlet elemeire a modell által adott kategóriákat és ezt összevetjük a helyes kategorizálással, és valamilyen módon kvantitatívan értékeljük a teljesítményt

2.6 Alkalmazás a magyar nyelvre

A fenti módszercsalád sikere angol nyelvű szövegek osztályozására bizonyítja az alapvető redukciós ötlet helyességét: csak bizonyos kulcsfogalmak előfordulását vizsgálva a szövegekben - minden mást, a további szavakat és a sorrendiségi információkat eldobva - jó minőségű osztályozót készíthetünk.

Magyar nyelvű szövegek esetén azonban a fenti technikák változatlan formában történő használata lényegesen rosszabb eredményre vezet. Ez a jelenség könnyen megérthető. Ugyanis a klasszifikáció szempontjából egy bizonyos szó különböző toldalékolt alakjai lényegében ugyanazt az információt tartalmazhatják, ám mi teljesen független megjelenésekként kezeljük őket. Ezzel a helyes relevancia megállapítását is lehetetlenné tesszük, hiszen pl. könnyen előfordulhat, hogy egy releváns fogalom különbözőképpen ragozott alakjai közül egyik sem releváns (pl. egyszerűen a ritka előfordulásuk miatt). Másrészt ha egy szó több alakját is kiválasztjuk, akkor a kapott bináris vektoroknak különböző koordinátái azonos jelentésűek lesznek, feleslegesen nehezítve ezzel az SVM dolgát.

3 A magyar nyelvhez illesztés lehetőségei

Ebben a fejezetben általánosítjuk a fent leírt algoritmust, majd megadjuk a kapott általános algoritmus két, a magyar nyelv jellegéhez a klasszikus technikánál jobban illeszkedő speciális esetét.

3.1 Jellemző előfordulási modell

Vegyük észre, hogy a 2.5. szakasz lépései közül sehol sem használtuk ki, hogy amiknek az előfordulását vizsgáljuk a dokumentumokban, azok szavak. Valójában általánosítható a dokumentum ábrázolási modellünk a következő módon. Rögzítünk valahogy jellemzők egy (nem feltétlenül véges) halmazát úgy, hogy bármely szöveghez meghatározható legyen ennek egy véges részhalmaza, ami az előforduló jellemzőknek felel meg, és ezzel a részhalmazzal fogjuk reprezentálni a dokumentumainkat. (A jellemzők alaphalmazának elemei az eddigiekben az alfanumerikus karaktersorozatok voltak, és egy dokumentumhoz a benne elválasztó karakterek között előforduló sorozatok halmaza tartozott.)

A fenti reprezentációval adott tanító dokumentumkészlet segítségével kiválaszthatjuk a releváns jellemzőket, és azokból épített bináris vektorokat osztályozhatunk, azaz tulajdonképpen módosítás nélkül alkalmazhatjuk a fenti technikákat. Ennek az általánosabb reprezentációnak az előnye, hogy a jellemzők ügyes megválasztásával lényegesen javíthatunk az eredményeken.

3.2 Szótövek mint jellemző

A 2.6. részben leírt probléma kézenfekvő megoldása, hogy az előforduló szavakat a további feldolgozás előtt szótövesítjük. Tehát a vizsgált jellemzők szótő előfordulások lesznek. Erre a célra a JMorph morfológiai motort használtuk (lásd [8]).

Fontos tervezési kérdés, hogy pontosan mit is tekintsünk szótőnek, hiszen nem érdemes minden esetben az összes toldaléktól megszabadulni (jó példa erre az egészség szó). Az általunk választott megoldás az, hogy az összes jeltől és ragtól megválunk, és addig töröljük a képzőket, amíg egy már önállóan is a szótárunkban előforduló alakot nem kapunk. Ez egy jó heurisztika a megállási pont megválasztására, de persze nem adja minden esetben az optimális megoldást.

A technika hátránya a viszonylag nagy apparátus igénye. Egy morfológiai elemző önmagában bonyolultabb mint a rendszer többi része együttvéve, nem is beszélve a mögötte álló morfológiai erőforrások elkészítésének munkaigényéről (szótár, szabálykészletek). A problémát tovább súlyosbítja, hogy ha új nyelvre akarunk osztályozót készíteni, az elemző készítését lényegében nulláról kell kezdenünk.

A szótövesítés futási idő szempontjából is költséges művelet, egy komoly keresési feladat. Érdemes itt megjegyezni, hogy különösen a sikertelen elemzések tartanak sokáig, hiszen jó keresési stratégiával a legjobbnak tűnő felbontás (ha létezik) az esetek túlnyomó többségében a teljes keresési tér bejárása nélkül is megtalálható, míg negatív válasz csak az összes eshetőség végigpróbálása után adható. Tehát az ismeretlen ill. elírt szavak elemzése sokáig is tart és persze végül nem is kapunk használható eredményt. (Azon szavakat, amelyekkel az elemző nem tud mit kezdeni kísérleteinkben önálló jellemzőknek tekintjük.)

3.3 N-gramm előfordulások mint jellemzők

Másodikként egy nyelvészeti tudást nem használó, igen egyszerű módszert javaslunk. Legyenek a jellemzők egyszerűen a szövegben előforduló egymás utáni betű n-esek. Variációs lehetőség, hogy a szóhatárokon átnyúló n-grammokat eldobjuk vagy megtartjuk. A „Hideg van.” mondatban előforduló betű hármasok pl. az első esetben {'hid', 'ide', 'deg', 'van'} a másodikban pedig {'hid', 'ide', 'deg', 'eg ', 'g v', ' va', 'van'}. Mi előzetes kísérletek alapján az első változatot használtuk.

Ezzel a technikával láthatóan rengeteg teljesen értelmetlen és véletlenszerű jellemzőt kapunk, ám a válogatási lépés ezeket automatikusan eltávolítja, így az osztályozáshoz már csak a tényleg sokat mondó n-grammok maradnak meg. Nagy előny, hogy a technika teljesen nyelvfüggetlen, csak a tanítási fázist kell újra elvégezni ha más nyelvű szövegeket szeretnénk osztályozni.

Az n-gramm alapú klasszifikáció további előnye, hogy várhatóan robosztusabban viselkedik zajjal szemben (pl. elgépelések, helyesírási hibák). Egy elírt szóra a morfológiai elemző szinte biztosan hibás eredményt ad, míg egy tipikus darabja jóval nagyobb eséllyel érintetlenül megtalálható. Olyankor is alkalmazhatók az n-grammok, amikor a szóhatárok nem ismertek, pl. beszédfelismerő rendszerek kimenetének osztályozása.

4 Elvégzett kísérletek, eredmények

A kísérletek vezérfonala, hogy a klassztterező metodika elemeinek nagy részét rögzítve megvizsgáljuk, hogy a jellemzők halmazának megválasztása milyen hatással van az osztályozás minőségére.

4.1 A teljes kísérlet sorozat során rögzített választások

1. A tanító- és tesztkorpusz szétválasztása minden egyes kísérletben függetlenül és véletlenszerűen történt, 70-30 arányban (a dokumentumok 70% került a tanítókorpuszba).
2. A relevancia megállapítását az információs nyereség függvény segítségével végeztük. A kiválasztandó szavak *k* számát egy logaritmikus skála mentén változtattuk, tehát az egyes technikák minőségét egy függvénnyel jellemeztük, melynek változója a kiválogatott jellemzők száma. A rögzített k értékek: 10, 16, 27, 46, 77, 129, 216, 362, 604, 1010.
3. Az osztályozást lineáris SVM-mel végeztük, a *c* paraméter 1.0 értékével. Ehhez a libsvm (lásd [2]) nevű szabadon használható implementációt használtuk.
4. Az eredmények kvantitatív értékelése az f-measure értékelő függvénnyel történt (lásd pl. [10]).

4.2 Használt korpuszok

4.2.1 Az Index hírportálról letöltött cikkek

A kizárólag magyar nyelvű kísérletekhez az Index [4] nevű hírportál archívumából letöltött 2100 cikket használtuk. A cikkek természetes kategorizálását adja a rovat neve. Az alábbi hét rovatból szerepelnek cikkek a korpuszban, rovatonként 300 cikkel: bulvár, gazdaság, külföld, kult, sport, tech-tudomány.

4.2.2 A Hunglish korpusz

A Hunglish korpusz a Budapesti Műszaki Egyetem Média Oktató és Kutató Központja és a Magyar Tudományos Akadémia Nyelvtudományi Intézet Korpusznyelvészeti Osztálya által készített több mint 50 millió szövegszót és 2 millió mondatot tartalmazó, mondatszinten illesztett, magyar-angol párhuzamos korpusz [9]. A magyar és angol nyelvű szövegek osztályozását összehasonlítandó ebből a korpuszból választottunk ki jogi és irodalmi szövegeket. Minden egyes osztályozandó dokumentumot a korpusz 50 egymást követő mondatból állítottuk össze, ilyen dokumentumból ezret-ezret készítettünk el a tesztjeinkhez. Az egyes dokumentumok kategóriája "law" ill. "lit" volt attól függően,. hogy jogi vagy irodalmi szövegrészletről volt-e szó.

4.3 Szó alapú osztályozás hatékonyságának összehasonlítása angol és magyar nyelvű szövegekre

Az elsőként ismertetendő kísérletünk célja az volt, hogy megvizsgáljuk, hogy a klaszszifikációs feladat szó alapú megoldásának hatékonyságáért mennyiben felelős a dokumentumok nyelve. Erre a célra tökéletesen alkalmas volt a Hunglish korpusz, hiszen segítségével pontosan ugyanazon szövegek magyar és angol nyelvű változatain végezhettünk osztályozást. A feladat a jogi és irodalmi szövegek elválasztása volt, a kiválasztott szavak számának függvényében a kapott eredményeket az alábbi grafikon foglalja össze.

A várakozásoknak megfelelően az angol nyelvű dokumentumok elkülönítése jobb eredményt adott kevés releváns szót kiválogatva. Meglepő, hogy nagy jellemzőszámok esetén a magyar nyelvű klassztterezés egy árnyalattal sikeresebb. A vizsgált feladat meglehetősen könnyű, és nem tipikus abból a szempontból, hogy itt nem csak (sőt nem elsősorban) a dokumentumok témája lehet az elkülönítés alapja, hanem azok stílusa, szóhasználata. Így a meglepő jelenség magyarázata lehet az, hogy a magyarban nagyobb különbség van a jogi és a szépirodalmi stílus között.

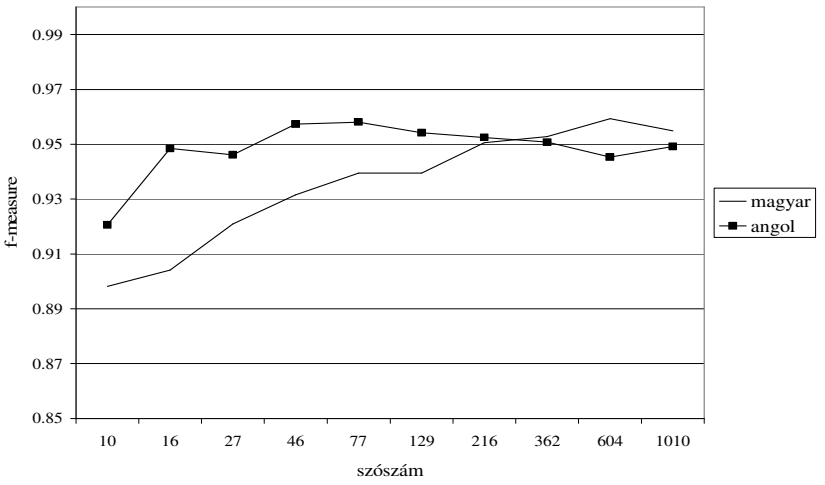

Fig. 1. A Hunglish korpusz jogi dokumentumainak elkülönítése az irodalmiaktól szavak előfordulásai alapján, magyar és angol nyelven

A várakozásoknak megfelelően az angol nyelvű dokumentumok elkülönítése jobb eredményt adott kevés releváns szót kiválogatva. Meglepő, hogy nagy jellemzőszámok esetén a magyar nyelvű klasszterezés egy árnyalattal sikeresebb. A vizsgált feladat meglehetősen könnyű, és nem tipikus abból a szempontból, hogy itt nem csak (sőt nem elsősorban) a dokumentumok témája lehet az elkülönítés alapja, hanem azok stílusa, szóhasználata. Így a meglepő jelenség magyarázata lehet az, hogy a magyarban nagyobb különbség van a jogi és a szépirodalmi stílus között.

A továbbiakban érdemes lenne egy tipikusabbnak mondható topic felismerési feladatot is kipróbálni a korpuszon, de ehhez szükség van egy jó minőségű tanító korpuszra, melynek előállítására ezen cikk megszületéséig nem volt lehetőségünk.

4.4 Az n érték választásának hatása n-grammos klasszifikáció esetén

Ebben az előkészítő jellegű mérésben arra kerestük a választ, hogy n értékének megválasztása hogyan befolyásolja az n-gramm alapú klasszifikáció eredményét. Megvizsgáltuk különböző n-ekre a jellemzőszám függvényében kapott osztályozási pontosságot, eredményeink az 2. ábrán és 3. ábrán láthatóak.

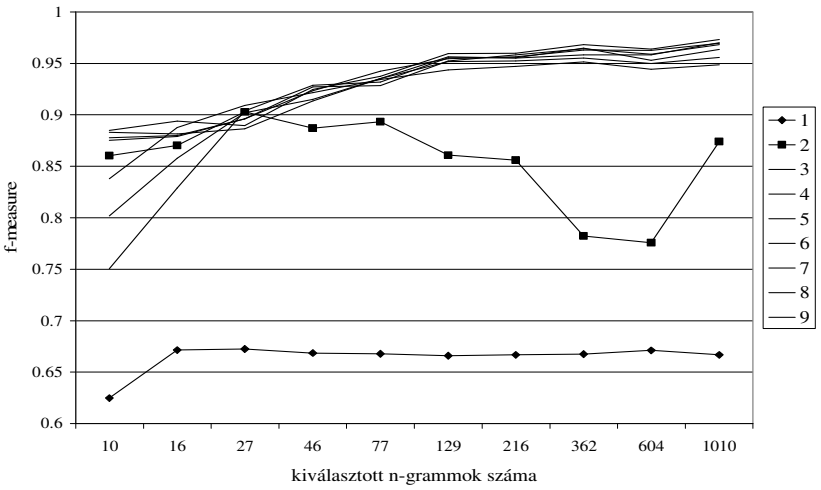

Fig. 2. Az Index Sport rovatába tartozó cikkek felismerése n-grammok (n=1, 2, 3, 4, 5, 6, 7, 8, 9) előfordulásai alapján, a használt releváns jellemzők számának függvényében

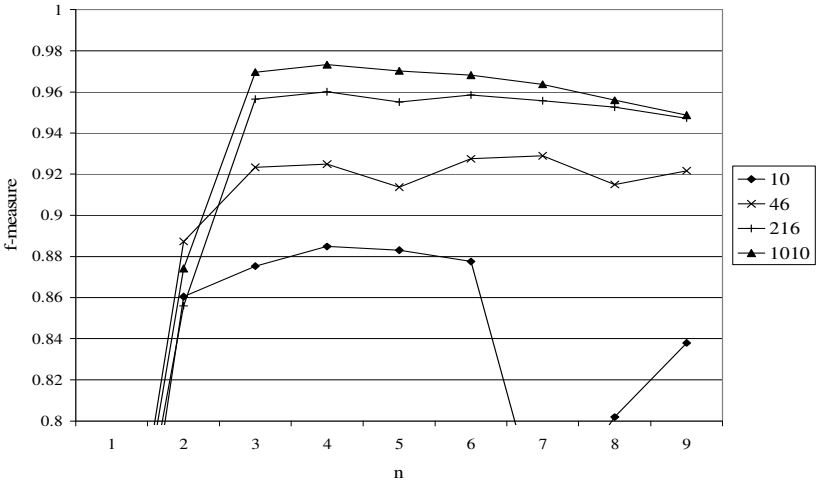

Fig. 3. Az Index Sport rovatába tartozó cikkek felismerése n-grammok előfordulásai alapján, 10, 46, 216 ill. 1010 legrelevánsabb n-gramm alapján osztályozva *n* függvényében

Az 1-gramm alapú osztályozás a várakozásoknak megfelelően használhatatlan eredményt ad, a 2-gramm már lényegesen jobb, de még mindig sokkal gyengébb a 3-gramm alapúnál, ám n=3-tól kezdve n-et tovább növelve már nem javul lényegesen az eredmény. A magasabb n-ekre a teljesítmény alakulása a 3. ábrán jobban megfigyelhető. A fenti és további hasonló kísérletek alapján végül is az n=4 választás mellett döntöttünk.

4.5 A jellemző választás hatása a magyar nyelvű szövegek osztályozására

Ebben a – cikk fő mondanivalóját alátámasztó – kísérletben az Indexes korpuszon végeztünk szövegklasszifikációt különböző jellemzőhalmaz választások mellett. A kipróbált jellemzőhalmazok a fentiekben részletesen ismertetett szavak halmaza (a klasszikus megoldás), szótövek halmaza (az előzőből szótövesítéssel kapjuk) ill. 4-grammok halmaza. A feladat a sport kategóriába tartozó cikkek elválasztása volt az összes többitől. A kapott eredményeket lásd a 4. ábrán!

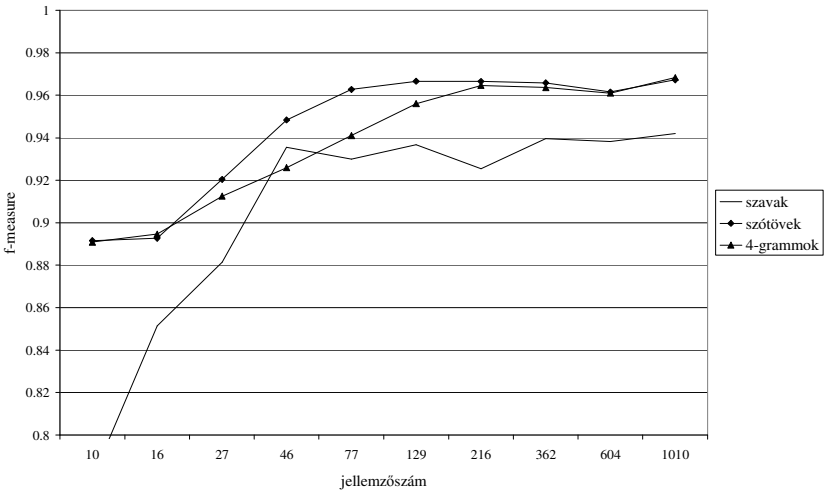

Fig. 4. Az Index Sport rovatába tartozó cikkek felismerése szó, szótő ill. n-grammok (n=4) előfordulásai alapján, a használt releváns jellemzők számának függvényében

Jól megfigyelhető, hogy a legjobb megoldást a szótövesítés adja, de a 4-gramm alapú technika teljesítménye is legfeljebb egy százalékkal marad el, közepes jellemzőszámok mellett. Nagy és kicsi jellemzőszámok mellett a két technika közel azonos teljesítményt ad.

4.6 Zajos dokumentumok vizsgálata

Ezzel a kísérlettel azt a feltevésünket próbáltuk ellenőrizni, hogy az n-gramm alapú klasszifikáció a másik két módszernél lényegesen robosztusabb hibás szövegek esetén. A szövegekben előforduló hibák szimulálására minden egyes karaktert egy adott valószínűséggel egy véletlenszerűen kiválasztott másikra cserélünk. Itt rögzítjük a jellemzőszámot, a 216 legrelevánsabb jellemzőt használjuk (ezt a választást az előző rész eredményei támasztják alá). A hibavalószínűség függvényében a három módszer által adott klasszifikáció minőségét az alábbi ábra foglalja össze.

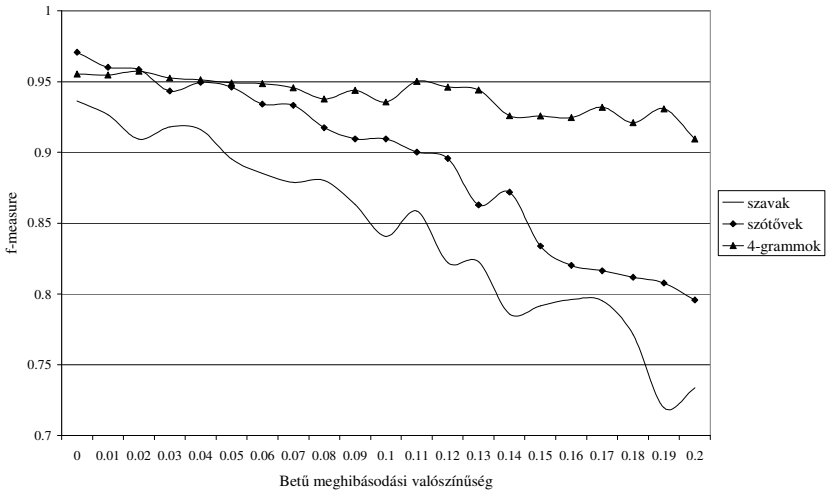

Fig. 5. Az Index Sport rovatába tartozó cikkek felismerése szó, szótő ill. n-grammok (n=4) előfordulásai alapján, a hozzáadott zaj mértékének függvényében

Az n-grammos módszer eredményei a zaj függvényében lényegesen lassabban romlanak, mint a másik két módszer esetében. Még 20%-os hiba esetén is kevesebb mint 5%-kal csökken az osztályozási hatékonyság. Ez fontos lehet olyan esetekben, amikor a szöveg valamilyen (pl. beszéd vagy írás) felismerés eredménye, vagy egyéb okból (pl. e-mail) zajos.

5 Következtetések

A magyar nyelvű szövegek klasszifikációjában komoly segítséget jelent, ha valamilyen módon megpróbálunk a különböző szóalakok közös kezelésére lehetőséget biztosítani. Erre megfelelő nyelvészeti apparátus birtokában a legjobbnak bizonyuló megoldást a szótövesítés adja. Azonban a sokkal kevesebb fejlesztési és futási időt igénylő n-gramm alapú osztályozók teljesítménye alig marad el a szótő alapú osztályozásétól.

A jelenség azzal magyarázható, hogy a betű hármasok, betű négyesek egy jelentős része már tipikusan egyetlen szó különböző szóalakjaira jellemző, így jó eszköz ezek összefogására, együtt kezelésére. Bizonyos n-grammok persze nem ilyenek, pl. a tipikus toldalékokhoz tartozó betűsorozatok, ám ezektől automatikusan megszabadulunk a relevancia szerinti válogatás során.

Ha a készítendő osztályozót várhatóan zajos dokumentumokra fogjuk alkalmazni, vagy ha nem ismerjük a szóhatárokat, akkor az n-gramm alapú osztályozás nem csak olcsóbb, de jobb megoldást is ad.

6 Köszönetnyilvánítás

A szerzők köszönetüket fejezik ki Gyepesi Györgynek és Varga Dánielnek a JMorph morfológiai elemzővel ill. a Hunglish korpusszal kapcsolatos értékes segítségükért. A jelen cikkben ismertetett kutatást részben a Kutatás-fejlesztési Pályázati és Kutatáshasznosítási Iroda GVOP-3.1.1.-2004-05-0363/3.0 számú Orvosi szak szövegek interaktív tartalomelemzése elektronikus kórlapok kitöltésére című pályázata támogatta.

Bibliográfia

1. Cavnar, W. B., Trenkle, J. M.: N-Gram-Based Text Categorization. In Proceedings of SDAIR-94 (1994) 161–175
2. Fan, R.-E., Chen, P.-H., Lin, C.-J.: Working set selection using the second order information for training SVM. Technical report, Department of Computer Science, National Taiwan University (2005), http://www.csie.ntu.edu.tw/~cjlin/libsvm/
3. Forman, G.: Choose Your Words Carefully: An Empirical Study of Feature Selection Metrics for Text Classification In Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery. Lecturre Notes in Computer Science, Vol. 2431. Springer-Verlag, London UK (2002) 150–162
4. Index.hu hírportál. http://www.index.hu
5. Kornai A., Krellenstein M., Mulligan M., Twomey D.,Veress F., Wysoker A.: Classifying the Hungarian web. In Copestake and Hajic (eds): Proceedings of EACL 2003 203–210
6. Kwok, J. T.: Automated text categorization using support vector machine. In Proceedings of ICONIP'98, 5th International Conference on Neural Information Processing, Kitakyushu, Japan (1998) 347–351
7. Langdon, W. B.: Natural Language Text Classification and Filtering with Trigrams and Evolutionary NN Classifiers. In Darrell Whitley (ed): Late Breaking Papers at the 2000 Genetic and Evolutionary Computation Conference, Las Vegas, Nevada, USA (2000) 210–217
8. Trón, V., Németh, L., Halácsy, P., Kornai, A., Gyepesi, Gy., Varga D.: Hunmorph: open source word analysis. In Proceedings of ACL Software Workshop (2005) 77–85
9. Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V.: Parallel corpora for medium density languages. Proeedings of RANLP 2005, to appear
10. Yang, Y.: An Evaluation of Statistical Approaches to Text Categorization. Int. J. Information Retrieval Vol. 1/1-2 (1999) 69–90